

# Terminological and language resources for developing a virtual patient dialogue system in Spanish

## *Recursos lingüísticos y terminológicos para desarrollar un sistema de diálogo con un paciente virtual en español*

Leonardo Campillos-Llanos<sup>1, 2</sup>, Catherine Thomas<sup>1, 2</sup>, Éric Bilinski<sup>1</sup>,  
Pierre Zweigenbaum<sup>1</sup>, Sophie Rosset<sup>1</sup>

<sup>1</sup>LIMSI - CNRS

<sup>2</sup>SATT Paris Saclay

{campillos, bilinski, pz, rosset}@limsi.fr

**Abstract:** This article reports the terminology and language resources used to develop a dialogue system for the medical domain in Spanish. The system simulates the medical history-taking step with a virtual patient, and is aimed mainly at medical students, who can train their communication and anamnesis skills in e-learning contexts. The conversational agent was first developed in French, and only the French system was evaluated by potential end-users. Herein, we provide an overall description of the system with a special focus on the Spanish components.

**Keywords:** Sistemas de diálogo, Terminología, PLN en Medicina

**Resumen:** Este artículo presenta los recursos lingüísticos y terminológicos empleados para desarrollar un sistema de diálogo de dominio médico en español. El sistema simula la etapa de entrevista clínica con un paciente virtual, y se destina principalmente a estudiantes de Medicina, quienes pueden entrenar sus destrezas comunicativas y de anamnesis en contextos de enseñanza a distancia. El agente conversacional se desarrolló principalmente en francés, y solo la versión francesa fue evaluada por potenciales usuarios. En esta contribución, describimos de manera global el sistema, con un foco especial en los componentes para la lengua española.

**Palabras clave:** Dialogue systems, Terminology, Medical NLP

## 1 Introduction

Dialogue systems for health applications are a vibrant field of research. Literature reviews (López-Cózar et al., 2014; Laranjo et al., 2018; Montenegro, da Costa, and da Rosa Righi, 2019) show how applications are progressively widening their scope from telemedicine, patient follow-up and counselling, to virtual assistants for clinicians to get information from databases. Conversational agents are also used to interact with avatars in educational software for training healthcare professionals by means of virtual patients (VP) (Rombauts, 2014).

In a collaborative project with companies,<sup>1</sup> we collaborated in the creation of a virtual patient e-learning software for medical doctors. The software allows medical instructors to define a patient profile—with

its health problems and medical history—so that medical students can practice their diagnosis skills. Our contribution was to create a dialogue system for the medical history taking step: medical doctors can train their anamnesis skills through a conversational agent, which was developed in French, English and Spanish. Figure 1 shows a sample of dialogue with the Spanish system.

Creating such a dialogue system brought up several challenges, namely: 1) To manage adequately the terminological variation in a domain of large vocabulary; 2) To provide reliable answers—correct according to the patient record—and realistic replies—to be expected from a patient perspective. In former work (Campillos-Llanos et al., 2019), we already developed on those aspects and reported the system architecture. Herein, we summarize its overall functioning (§2), explain the adaptation of the French version to

<sup>1</sup><https://pvdial.limsi.fr/>

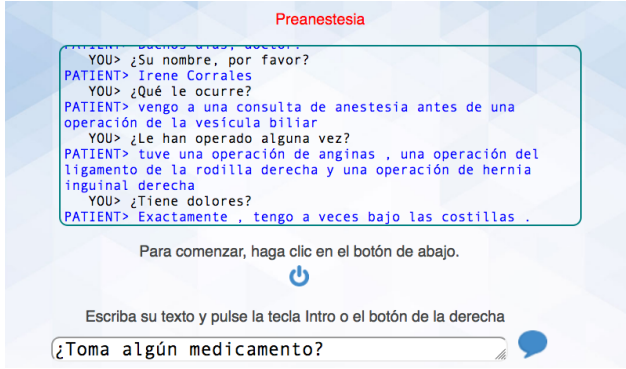


Figure 1: Sample of dialogue in Spanish

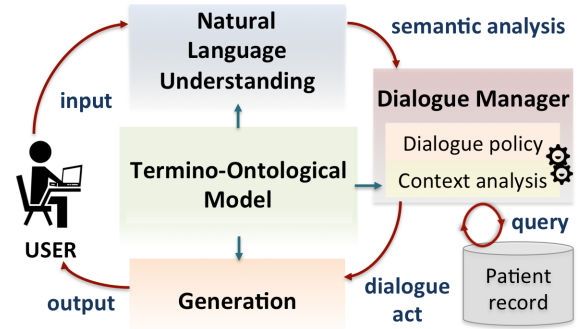


Figure 2: Schema of the dialogue system

the Spanish language (§3), and we describe the resources of the Spanish version (§4-§6). We conclude by reporting the current state of the system and future work (§7).

## 2 Dialogue System Architecture

The system is made up of different units (Figure 2), integrated in an infrastructure for dialogue systems (Rosset et al., 2005):

- A **Natural Language Understanding (NLU)** module: performs linguistic and semantic processing of user’s input.
- A **Dialogue Manager**: determines the dialogue policy, processes the information state at each dialogue move (according to the previous dialogue history) and queries the virtual patient record.
- A **Linguistic and Termino-ontological model**: manages the linguistic and terminological variation of the dialogue task.
- A **Generation module**: replies to the user by means of templates to be filled with the contents of the VP record.

## 3 Overall procedure of adaptation from French to Spanish

The dialogue manager did not need any language-specific adaptation, but the other components needed language resources to cope with the processing steps in the Spanish language. The NLU module makes use of a general dictionary, and domain lists, which were collected following the same procedures as those applied for the French components. We extracted medical terms from the Unified Medical Language System® (hereafter, UMLS) according to semantic contents needed for the dialogue task; we

summarize them in §4. We also translated lists from French to Spanish when no data could be extracted from domain terminologies; e.g. lay terms expressed in patient-language (e.g. *tener colesterol* instead of *hipercolesterolemia*); or general, out-of-domain lists (e.g. expressions of frequency or duration). Translation was enhanced with semi-automatic procedures and manually revised for quality check. Regular expressions and rules were translated manually.

As for the Termino-Ontological model, we collected equivalent resources to those needed in the French version (see §5). The Generation module needed both: 1) Collecting lexical resources from existing dictionaries and terminologies in Spanish; and 2) Translating manually the templates for generating the replies to be filled with the contents of the patient record. Finally, to build a small set of VP records for development, we translated from French 13 cases used for evaluating the French system. The translation of French to Spanish was rather feasible and fast (around 6 months). Both languages are close and have available terminological or language resources of similar type and size for the task.

## 4 Resources for Natural Language Understanding

This stage involves the following steps:

- **Linguistic preprocessing**: tokenization and Part-of-Speech (PoS) tagging with TreeTagger (Schmid, 1995). We use a dictionary of 627,099 entries for lemmatizing general and domain terms.
- **Spelling correction**: a spell-checker corrects the input or asks the user to confirm the correction of a misspelling.

- **Semantic annotation:** we use rules developed in Wmatch (Galibert, 2009), an engine for applying regular expressions using gazetteers. We created 149 NLU labels for annotating entity types (e.g. *symptom*), dialogue acts (e.g. *greeting*) and question types (e.g. *Qwhere*).

For each of the 149 labels, we formalized a grammar by means of regular expressions, rules of sentence structures and gazetteers for Named Entity Recognition. The NLU labels cover the range of topics of the medical history taking: patient’s demographic data (e.g. name and age), medical history (e.g. diseases or surgeries), symptoms, medication intake, as well as lifestyle and social behavior. Some rules manage conversational acts (e.g. greetings), but out-of-the-task and out-of-domain acts are not covered (e.g. a question such as *What is your favourite film?*).

Lists of entities were extracted from the UMLS (Bodenreider, 2004), vs. 2017AA; or compiled manually. Lists of medical drugs come from the Spanish Drug Effect database (Segura-Bedmar et al., 2015) and the Agencia Española de Medicamentos y Productos Sanitarios.<sup>2</sup>

## 5 Resources for the Linguistic and Termino-Ontological Model

This component gathers structured thesauri for managing linguistic and terminological variation: 1) To match input terms against the content of the patient record (*concept normalization*); and 2) To favour replies using non-technical terms. To query the patient record, entities detected in the user’s input are lemmatized and lowercased.

**Linguistic variation** requires managing inflectional and derivational variants, roots/affixes/stems, and general synonyms. **Inflectional variants** (e.g. *estornuda* ↔ *estornudo*) come from a lexicon gathering 45,763 word entries and 475,652 forms. **Derivational variants** (e.g. *fiebre* ↔ *febril*) and general **synonyms** (e.g. *andar* ↔ *caminar*) were obtained from the Spanish EuroWordNet in the Multilingual Central Repository 3.0 (Gonzalez, Laparra, and Rigau, 2012): for derivation, we used relations *RELATED\_TO*, *IS\_DERIVED* and *PERTAINS\_TO*; and for synonymy, the relation *SYNONYM*. **Roots, stems and affixes**

were translated from the lexicon used in the French version of the dialogue system, and also from English roots in the Specialist Lexicon (Browne, McCray, and Srinivasan, 2000).

To cater for **terminological variation** and map domain terms, we used the UMLS Metathesaurus, which makes it possible to map medical terms referring to the same concept, identified by a Concept Unique Identifier (CUI). Terms were collected from semantic type T121 (Pharmacological substance, e.g. *calmante* ↔ *analgésico*, C0002771, 2.5.1) and 3 semantic groups: ANAT, for anatomic entities (e.g. *vientre* ↔ *abdomen*, C0022658); DISO, for syndromes and disorders (e.g. *enfermedad de riñón* ↔ *nefropatía*, C0022658); and PROC, for procedures (e.g. *operación de hernia* ↔ *herniorrafía*, C0019328). We clustered in *ad hoc* lists domain synonym terms not found in the UMLS: e.g. symptom verbs and nouns (*adelgazar* ↔ *adelgazamiento*).

Lastly, we gathered **semantic relations** between CUIs from the UMLS Semantic Network (McCray, 1989). This helps the system to match input terms to concepts in the VP record beyond synonymy, when a hierarchical relation exists. For example, *CHILD\_OF* relations may match a general entity to a specific entity (e.g. *enfermedad cardiovascular* ↔ *hipertensión*). We also use associative relations to link procedures and disorders (e.g. *herniorrafía* ↔ *hernia*), procedures and anatomic entities (e.g. *apendicectomía* ↔ *apéndice*), and pathologies and physiological entities (e.g. *vómito* ↔ *digestión*). These relations hold between concepts (not terms), hence they are language-independent.

## 6 Resources for Generation

This step involves three subtasks:

- **Gender/number agreement:** information extracted from the VP record is output through predefined templates, where nouns or adjectives need to agree. To generate the number and gender variant forms of the words to be output, we use linguistic information from DELAS-type dictionaries (Courtois, 1990).
- **Change of the content in the record to the patient’s viewpoint:** in the generation step, data expressed in 3rd person is changed to the 1st person (e.g. *[el paciente] tose* → *[yo] toso*); we use a

<sup>2</sup>[www.aemps.gob.es/](http://www.aemps.gob.es/)

table with correspondences between 3rd and 1st person verb forms.

- **Reply with non-technical terms:** we favour system replies according to a patient's perspective (e.g. *anginas* instead of *amígdalas*); to do so, we use lists of equivalent terms.

## 7 Discussion and Conclusion

The tool is one of the few dialogue systems in Spanish for such a task in the medical domain. The system aims at filling the need for improving doctors interaction skills by means of an e-learning tool as a complement to real clinical practice. Developing the system required us to tackle challenges of the task and domain—e.g. managing terminological variation, generate realistic output, favouring correctness in system replies. All of those aspects revitalize an area of research with problems still not solved satisfactorily.

Our experience shows that adapting the system to a close language is feasible with the methods explained. This could spark interest in adapting the system to other Romance languages, provided that similar language and terminological resources exist for the task.

To date, only the French version of the system was evaluated by potential end-users (i.e. medical students and doctors,  $n=39$ ). We conducted a quantitative evaluation—through the analysis of users' dialogue logs—and qualitative user evaluation—by means of 5-point Likert-scale questionnaires, collected with an on-line form which is also available for assessing the Spanish system. We aim at collecting more data to improve the Spanish version and evaluate it with medical students or residents. For that purpose, we prepared 13 different patient records by translating some cases used in the evaluation of the French system. The system can be tested and evaluated on-line.<sup>3</sup>

## Acknowledgements

The SATT Paris Saclay funded this work.

## References

- Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.
- Browne, A. C., A. T. McCray, and S. Srinivasan. 2000. The Specialist Lexicon. *National Library of Medicine Technical Reports*, pages 18–21.
- Campillos-Llanos, L., C. Thomas, E. Bilinski, P. Zweigenbaum, and S. Rosset. 2019. Designing a Virtual Patient Dialogue System Based on Terminology-rich Resources: Challenges and Evaluation. *Natural Language Engineering*. <https://doi.org/10.1017/S1351324919000329>.
- Courtois, B. 1990. Un système de dictionnaires électroniques pour les mots simples du français. *Lang. française*, 87(1):11–22.
- Galibert, O. 2009. *Approaches and methodologies for automatic Question-Answering in an open-domain, interactive setup*. Phd dissertation, Université Paris Sud, June.
- Gonzalez, A., E. Laparra, and G. Rigau. 2012. Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.
- Laranjo, L. et al. 2018. Conversational agents in healthcare: a systematic review. *J. Am. Med. Inf. Assoc.*, 25(9):1248–1258.
- López-Cózar, R., Z. Callejas, D. Griol, and J. F. Quesada. 2014. Review of spoken dialogue systems. *Loquens*, 1(2):012.
- McCray, A. T. 1989. The UMLS Semantic Network. In *Proc. Symp. on Comp. Applic. Medical Care*, pages 503–507. AMIA.
- Montenegro, J. L. Z., C. A. da Costa, and R. da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications*.
- Rombauts, N. 2014. *Patients virtuels: pédagogie, état de l'art et développement du simulateur Alphadiag*. Phd dissertation, Faculty of Medicine, Lyon, France.
- Rosset, S., O. Galibert, G. Illouz, and A. Max. 2005. Interaction et recherche d'information: le projet Ritel. *Traitement automatique des langues*, 46(3):155–179.
- Schmid, H. 1995. Treetagger—a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.
- Segura-Bedmar, I., P. Martínez, R. Revert, and J. Moreno. 2015. Exploring Spanish health social media for detecting drug effects. In *BMC Medical Informatics and Decision Making*, volume 15, page S6.

<sup>3</sup>[www.audiosurf.net/pg\\_2018/select\\_case.php](http://www.audiosurf.net/pg_2018/select_case.php)